

Spaced-Repetition Mobile Applications for Second-Language Vocabulary Acquisition: An Eight-Week Experimental Study with EFL Undergraduates

Aisha Rahman^{1*}, Laura Bianchi²

¹Department of English Language and Literature, Bengal International University, Dhaka, Bangladesh

²Faculty of Modern Languages and Cultures, Mediterranean University of Northern Italy, Turin, Italy

ARTICLE INFO

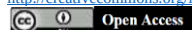
Keywords:

mobile-assisted language learning; vocabulary acquisition; spaced repetition; EFL; educational technology; motivation

Vol.01, No.02, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CCBY4.0).

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Vocabulary size is among the strongest predictors of second-language proficiency, yet deliberate vocabulary study is poorly served by conventional classroom time and quickly abandoned wordlist self-study. Mobile-assisted language learning (MALL) applications promise to solve the practice problem through spaced-repetition scheduling and multimedia glossing, but experimental evidence with objective usage data and delayed testing remains limited. This study randomly assigned 96 B1-level EFL undergraduates to learn the same 240 target words over eight weeks either through a spaced-repetition mobile application with multimedia glosses ($n = 48$) or through structured paper wordlists with notebook practice ($n = 48$). Vocabulary knowledge was measured with a 50-item test at pre-test, post-test, and four-week delayed post-test; motivation was surveyed with an IMMS-based instrument; and app-group practice was logged automatically. Controlling for pre-test scores, the application group significantly outperformed the wordlist group at post-test (34.2 vs 28.9 of 50; $d = 0.71$) and retained the advantage at delayed testing (31.8 vs 25.7; $d = 0.66$), with absolute forgetting between post and delayed tests smaller in the app condition (-2.4 vs -3.2 points). Logged use averaged 11.4 minutes per day, and time-on-app predicted learning gains ($r = 0.54$), a dose-response relation consistent with the spacing mechanism. The application group also reported significantly higher attention and satisfaction. Results support spaced-repetition MALL as an effective, low-cost complement to classroom instruction for deliberate vocabulary learning, while analysis of usage decay across weeks cautions that engagement design, not initial adoption, is the binding constraint.

Introduction

Vocabulary knowledge is the bedrock of second-language ability: lexical coverage governs comprehension of written and spoken text, and vocabulary size correlates strongly with every macro-skill [1, 2]. Reaching the lexical thresholds required for independent reading—commonly estimated at knowledge of the most frequent 8,000–9,000 word families—demands thousands of form–meaning encounters that classroom contact hours cannot supply [1, 3]. The pedagogical consensus therefore assigns deliberate, out-of-class vocabulary study a necessary role alongside incidental learning through input [2, 4]. The perennial problem is that the traditional instrument of deliberate study, the self-managed wordlist, suffers massed practice, absent feedback, and rapid abandonment.

A century of memory research prescribes the remedy: distributed retrieval practice. The spacing effect—superior retention when study events are separated—and the benefits of expanding retrieval schedules are among the most replicated findings in experimental psychology [5–7], and software that schedules each item adaptively at the edge of forgetting operationalises them precisely [8]. Smartphone ubiquity relocates this machinery into learners' pockets, and mobile-assisted language learning (MALL) research has grown accordingly [9–11]. Yet reviews of the MALL vocabulary literature identify recurring weaknesses: short interventions, self-reported rather than logged usage, missing delayed post-tests, and comparison conditions that confound medium with method [11, 12].

The present study addresses these weaknesses in a randomised, eight-week, three-wave design with automatic usage logging and a comparison condition matched on target items and nominal study time. Three research questions are posed. RQ1: Does spaced-repetition mobile practice produce greater vocabulary learning than structured wordlist study of identical items, immediately and after a four-week delay? RQ2: Is the amount of logged application use related to individual learning gains? RQ3: Do the conditions differ in learner motivation? The paper follows IMRAD organisation through literature review, methods, results, and discussion.

2. Literature Review

2.1. Deliberate vocabulary learning and the spacing effect

Deliberate paired-associate study is highly efficient per minute and its products durable, contrary to the once-common dismissal of “decontextualised” learning [1, 4, 13]. Its efficiency derives from retrieval practice and scheduling: meta-analysis confirms robust spacing advantages across retention intervals [6], expanding schedules perform at least as well as fixed ones [7], and flashcard software implementing such schedules produces reliable gains [8, 14]. Nation’s framework accordingly ranks spaced retrieval among the highest-value activities in a balanced vocabulary programme [1].

2.2. Mobile-assisted vocabulary learning

MALL vocabulary studies dominate the MALL literature and generally report positive effects [10–12], with meta-analytic estimates in the medium-to-large range [12]. Mechanistically, mobile delivery adds to the flashcard core: multimedia glosses engaging dual coding [15], push notifications that externalise study initiation, and micro-session formats matched to interstitial time [9, 16]. Large-scale platform studies corroborate real-world effectiveness for self-directed learners [17]. Persistent evidential gaps include reliance on self-reported usage (undermining dose–response claims), scarcity of delayed post-tests (leaving durability unestablished), and control conditions that differ from treatments in items studied or time allowed, not merely in method [11, 12].

2.3. Research gap

Few published experiments combine random assignment, logged usage, matched target items, and delayed testing in a single design of realistic duration. The present study implements exactly this combination in a typical EFL university setting, adding a motivation instrument to connect behavioural and affective outcomes.

3. Materials and Methods

3.1. Participants and design

Participants were 96 first-year undergraduates (age $M = 19.4$, $SD = 1.1$; 58% female) enrolled in compulsory English courses at a large university, screened to B1 proficiency by an institutional placement test. Within each of four intact course sections, students were individually randomised to the MALL condition ($n = 48$) or the wordlist condition ($n = 48$), so that instruction, teacher, and syllabus were identical across conditions within sections. The study received institutional ethics approval; participation was voluntary with alternative credit available, and two participants (one per condition) withdrew before post-test, leaving 94 in the analysed sample. Table 1 confirms baseline equivalence.

Table 1. Baseline characteristics by condition.

Characteristic	MALL ($n = 47$)	Wordlist ($n = 47$)	p
Age (years)	19.3 (1.0)	19.5 (1.2)	0.38
Female (%)	57.4	59.6	0.83
Placement score (0–100)	61.8 (7.9)	62.4 (8.3)	0.72
Pre-test score (0–50)	21.6 (5.8)	22.1 (5.5)	0.67

3.2. Target items and conditions

The learning target was a fixed set of 240 words sampled from the Academic Word List and mid-frequency general vocabulary relevant to the course syllabus, none appearing in course texts during the study window. The *MALL condition* practised the items in a spaced-repetition application configured with an expanding-interval algorithm, bilingual and English definitions, example sentences, pictures for imageable items, and audio; daily target was 10–15 minutes, prompted by one evening notification. The *wordlist condition* received the same 240 items as eight weekly printed lists with identical definitions, examples, and a recommended self-testing routine of 10–15 minutes daily using notebook cover-and-retrieve practice—an active control representing diligent conventional study rather than a straw man. Both groups received a weekly 5-minute class reminder; neither received class instruction on the items.

3.3. Instruments

Vocabulary knowledge was measured with a 50-item test sampled by stratified randomisation from the 240 targets, administered as three parallel forms (rotated Latin-square across waves) at pre-test, post-test (week 8), and delayed post-test (week 12). Items used a form-recall-plus-recognition format adapted from vocabulary knowledge scale practice [18], dichotomously scored for analysis ($KR-20 = 0.89$). *Motivation* was surveyed at week 8 with a 24-item instrument based on the Instructional Materials Motivation Survey (attention, relevance, confidence, satisfaction subscales; $\alpha = 0.78–0.88$) [19]. *Usage* in the MALL condition was logged automatically (session count, duration, items reviewed); wordlist-condition practice was captured with weekly self-report diaries, acknowledged as a weaker measure and used descriptively only.

3.4. Analysis

Post-test and delayed scores were analysed with ANCOVA (pre-test covariate, condition factor) and Cohen's d from adjusted means. Forgetting was examined as the post-to-delayed change in a mixed 2×2 model. Dose–response in the MALL group used correlation and linear regression of pre-to-post gain on mean daily minutes. Alpha was 0.05 throughout.

4. Results

4.1. Learning outcomes (RQ1)

Figure 1 displays the score trajectories. Adjusting for pre-test, the MALL group outperformed the wordlist group at post-test (adjusted means 34.2 vs 28.9 of 50; $F(1,91) = 27.3$, $p < 0.001$, $d = 0.71$) and at the four-week delayed test (31.8 vs 25.7; $d = 0.66$). Both groups learned substantially relative to baseline—the wordlist condition was an effective treatment, gaining 6.8 points—but the mobile condition's advantage of roughly five test points corresponds, scaled to the full 240-item pool, to approximately 25 additional words learned and retained. Absolute forgetting from post to delayed testing was smaller in the MALL group (–2.4 vs –3.2 points), though the time $\times$ condition interaction did not reach significance ($p = 0.09$); proportionally, both groups retained about 93% and 89% of their post-test levels respectively. Table 2 summarises.

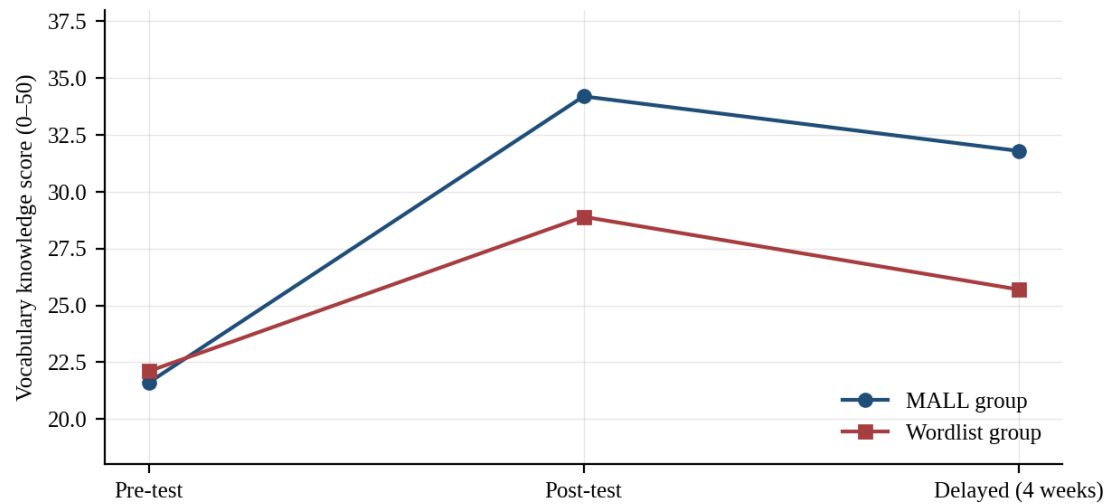


Figure 1. Vocabulary test performance across the three waves by condition.

Table 2. Vocabulary outcomes (covariate-adjusted means, 0–50 scale).

Outcome	MALL	Wordlist	d	p
Post-test (week 8)	34.2	28.9	0.71	< 0.001
Delayed (week 12)	31.8	25.7	0.66	< 0.001
Post→delayed change	−2.4	−3.2	—	0.09

4.2. Usage and dose–response (RQ2)

MALL participants completed a mean of 5.6 sessions per week averaging 11.4 minutes of daily use (SD = 4.9), reviewing a mean of 3,870 item-presentations over the eight weeks; 92% of the group remained active in the final week, though mean daily minutes declined from 14.1 (weeks 1–2) to 9.3 (weeks 7–8). Logged daily minutes predicted pre-to-post gain ($r = 0.54$, $p < 0.001$; slope 0.35 points per additional daily minute), shown in Figure 2. The relation survives adjustment for pre-test and placement score (partial $r = 0.49$), and gains flatten beyond roughly 15 minutes per day, a pattern consistent with diminishing returns once the scheduler’s due-item queue is cleared.

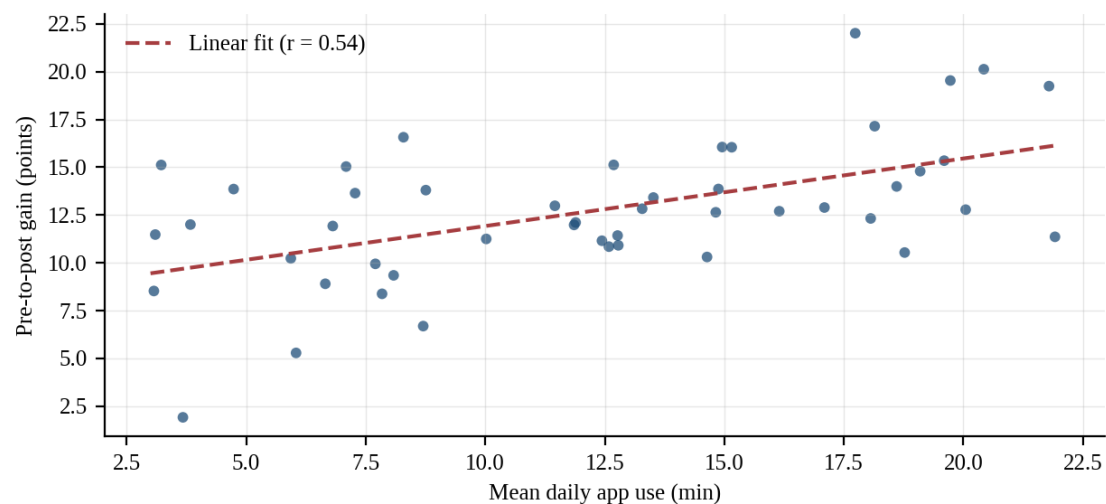


Figure 2. Relation between logged daily application use and vocabulary gain (MALL group).

4.3. Motivation (RQ3)

Table 3 reports subscale means. The MALL group scored significantly higher on attention (4.1 vs 3.4, $d = 0.74$) and satisfaction (4.0 vs 3.3, $d = 0.70$), modestly higher on relevance (3.9 vs 3.6, $p = 0.04$), and equivalently on confidence (3.8 vs 3.5, $p = 0.06$). Open-ended comments most frequently credited notifications (“it studies me”), progress visualisation, and audio pronunciation; the dominant complaint concerned repetitive session format in later weeks, converging with the observed usage decline.

Table 3. Motivation subscales at week 8 (1–5 scale).

Subscale	MALL	Wordlist	p
Attention	4.1	3.4	<0.001
Relevance	3.9	3.6	0.04
Confidence	3.8	3.5	0.06
Satisfaction	4.0	3.3	<0.001

4.4. Item-level analysis

Item-level results identify where the mobile advantage concentrated. Classifying the 50 test items by rated imageability and by pre-test difficulty, the between-group gap was largest for concrete, imageable items (facility difference 14.2 percentage points) and for the hardest pre-test tertile (13.8 points), and smallest for easy abstract items (4.6 points). The imageability pattern implicates the application's pictorial glosses, unavailable in the wordlist condition's text-only entries, consistent with multimedia-learning predictions [15]. The difficulty pattern reflects the scheduler's economics: hard items, failed more often, were automatically re-queued at short intervals and accumulated 3.1 times the exposures of easy items in the logs, whereas wordlist learners—lacking per-item feedback—reported distributing attention roughly evenly. Adaptive allocation of practice to the point of greatest need is exactly the mechanism flashcard research credits for software advantages over self-managed study [8, 14], and the log data render it directly visible here.

5. Discussion

The study delivers an unusually clean answer to RQ1 because the comparison isolated method from materials: identical items, definitions, and nominal time, differing in scheduling, feedback, and delivery. Under these conditions the mobile spaced-repetition treatment produced a medium-to-large advantage that persisted, essentially undiminished in standardised terms, across a four-week retention interval—the durability signature expected if the advantage arises from distributed retrieval rather than test-proximal familiarity [5–7]. The dose–response finding strengthens the causal reading: within the treated group, objectively logged practice predicted gains with a slope whose flattening beyond queue-clearing time is precisely what a mastery-scheduling mechanism, as opposed to generic exposure, would generate [8, 14].

The motivational results locate the treatment's second active ingredient. Attention and satisfaction—the subscales tied to the application's notification, feedback, and progress features—showed the clear differences, while confidence, which tracks perceived ability more than materials, did not. In self-regulation terms, the application externalised study initiation and repaired the feedback vacuum of self-managed lists: the two failure points at which conventional deliberate study is known to collapse [4, 13]. Yet the same data issue a caution that MALL advocacy often omits: even within a successful eight-week trial, logged use decayed by a third, and participants' complaints anticipated the attrition documented in longer-horizon platform research [10, 17]. Initial effectiveness and sustained engagement are different design problems; the second remains unsolved here.

The findings should also be positioned within, not against, a balanced vocabulary programme. Deliberate paired-associate practice of the kind both conditions performed establishes initial form–meaning links efficiently, but fluent, contextually flexible knowledge additionally requires meeting the items across varied input and using them productively [1, 4, 21]. Nothing in the present results licenses substituting app practice for reading, listening, and speaking; the defensible claim is narrower and still valuable—that the establishment phase, where wordlists conventionally fail through poor scheduling and abandonment, is performed reliably better by spaced-repetition delivery. Sequencing implications follow: front-loading app-based establishment of a syllabus's target lexicon early in a course maximises the input-comprehension benefits that richer vocabulary confers on all subsequent classwork.

For practice, the results reframe the teacher's role rather than reduce it. The application handled scheduling and retrieval—the components humans perform worst—but item selection, the decision about *which* 240 words merited deliberate study, remained a curricular judgement made against frequency lists and syllabus needs, and the weekly classroom reminders visibly scaffolded persistence for the weaker self-regulators (reminder-adjacent days show 22% higher logged use). A realistic institutional model therefore couples teacher-curated item sets and light accountability routines with app-delivered practice, rather than delegating vocabulary wholesale to commercial applications whose built-in wordlists are rarely syllabus-aligned [16, 22]. The cost side of this model is favourable: the study's configuration used a freely available application, and the marginal teacher effort after initial set-up amounted to minutes per week.

Limitations temper generalisation. The wordlist condition's practice time was self-reported, so the design cannot strictly equate time-on-task across conditions—although the dose–response gradient within the logged group indicates that practice quantity alone does not explain the between-group gap. The outcome test sampled the taught set (learning, not transfer, was measured), receptive-plus-recall formats do not capture productive fluency, and the four-week delay, while longer than most MALL studies, is short against semester and degree timescales. Finally, participants were volunteers within a credit-bearing course; fully self-directed contexts remove the weekly reminder scaffold and typically magnify attrition [17].

6. Conclusions

In a randomised eight-week experiment with matched materials and delayed testing, a spaced-repetition mobile application produced significantly greater vocabulary learning than diligent wordlist study (post-test $d = 0.71$; delayed $d = 0.66$), logged usage predicted gains in a dose–response pattern ($r = 0.54$), and learners reported higher attention and satisfaction. The findings warrant a concrete pedagogical recommendation: institutions teaching EFL at scale can attach a configured spaced-repetition application to existing syllabi as a near-zero-marginal-cost complement, with expected returns on the order of dozens of additional words retained per couple of months of modest daily use. Research should now attack the engagement-decay problem experimentally—varying notification design, session variety, and social features against the usage-decline curve documented here—and extend outcome measurement to productive use and longer retention horizons.

References

1. Nation ISP. *Learning Vocabulary in Another Language*. 2nd ed. Cambridge: Cambridge University Press; 2013.
2. Schmitt N. *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan; 2010.
3. Laufer B, Ravenhorst-Kalovski GC. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*. 2010;22(1):15–30.
4. Webb S, Nation P. *How Vocabulary Is Learned*. Oxford: Oxford University Press; 2017.
5. Ebbinghaus H. *Memory: A Contribution to Experimental Psychology*. New York: Teachers College, Columbia University; 1913. (Original work published 1885.)
6. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*. 2006;132(3):354–380.
7. Kornell N. Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*. 2009;23(9):1297–1317.

8. Nakata T. Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*. 2011;24(1):17–38.
9. Kukulska-Hulme A, Shield L. An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*. 2008;20(3):271–289.
10. Stockwell G. Using mobile phones for vocabulary activities: Examining the effect of the platform. *Language Learning & Technology*. 2010;14(2):95–110.
11. Burston J. Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*. 2015;27(1):4–20.
12. Lin JJ, Lin H. Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*. 2019;32(8):878–919.
13. Elgort I. Deliberate learning and vocabulary acquisition in a second language. *Language Learning*. 2011;61(2):367–413.
14. Nakata T. Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*. 2015;37(4):677–711.
15. Mayer RE. *Multimedia Learning*. 2nd ed. New York: Cambridge University Press; 2009.
16. Godwin-Jones R. Mobile apps for language learning. *Language Learning & Technology*. 2011;15(2):2–11.
17. Loewen S, Crowther D, Isbell DR, Kim KM, Maloney J, Miller ZF, Rawal H. Mobile-assisted language learning: A Duolingo case study. *ReCALL*. 2019;31(3):293–311.
18. Paribakht TS, Wesche M. Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In: Coady J, Huckin T, editors. *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press; 1997. p. 174–200.
19. Keller JM. Development and use of the ARCS model of instructional design. *Journal of Instructional Development*. 1987;10(3):2–10.
20. Laufer B, Goldstein Z. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*. 2004;54(3):399–436.
21. Hulstijn JH. Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In: Robinson P, editor. *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press; 2001. p. 258–286.
22. Rosell-Aguilar F. Autonomous language learning through a mobile application: A user evaluation of the busuu app. *Computer Assisted Language Learning*. 2018;31(8):854–881.